Regression Models - Peer Graded Assignment

Alejandro

27 Januar 2018

Introduction

In this report for Motor Trend, a magazine about the automobile industry we will be looking at a collection of car data and analyze their fuel consumption as miles per gallon (MPG). Here we will cover two main points

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions

Getting started

Load the package dependencies

The data set will be taken from the mtcars dataset in R.

```r
data("mtcars")
head(mtcars)
```

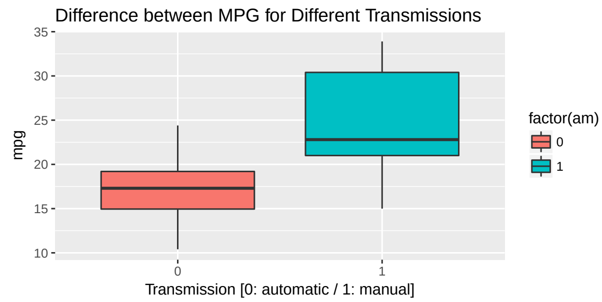
```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

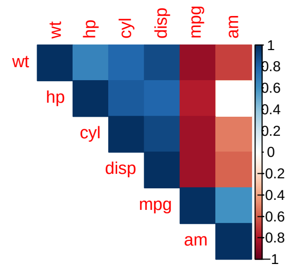
This dataset has 32 observations and 11 variables. These variables are:

- mpg Miles/(US) gallon
- cyl Number of cylinders
- disp Displacement (cu.in.)
- hp Gross horsepower
- drat Rear axle ratio
- wt Weight (1000 lbs)
- qsec 1/4 mile time
- vs V/S
- am Transmission (0 = automatic, 1 = manual)
- gear Number of forward gears
- carb Number of carburetors

Is an automatic or manual transmission better for MPG?

First we can do exploratory analysis on how the transmission affects the MPG.





In the figure on the left we see a boxplot comparing MPG for the two types of transmission. It is clear from the picture that manual transmissions have a better MPG as automatic. By doing a t.test we can reject the hypothesis that the transmission had no effect in mpg by having a p value of 0.001. By looking at the figure on the right we see that all variables with the exception of "am" have a negative correlation with the MPG.

Quantify the MPG difference between automatic and manual transmissions

Now we can fit a model to further analyze the effect of the transmission in the MPG. First we will start by having mpg as the output and just the transmission as the variable.

```r
mdl <- lm(mpg~factor(am), mydata)
summary(mdl)$coef[,1]
```

```
## (Intercept) factor(am)1
##   17.147368    7.244939
```

We see the in intercept being 17.15 and the coefficient 7.24. This means that that without including any other variables to our model we can expect an increase of **7.24** mpg when driving manual instead of automatic.

By adding more variables to our model the effect of course changes by explaining more of the variability of the model through more parameters. As an example we will fit again a linear model but this time adding two more variables: wt and cyl.

```r
mdl2 <- lm(mpg~factor(am) + wt + cyl,mydata)
summary(mdl2)$coef[,1]
```

```
## (Intercept) factor(am)1          wt         cyl
##  39.4179334   0.1764932  -3.1251422  -1.5102457
```

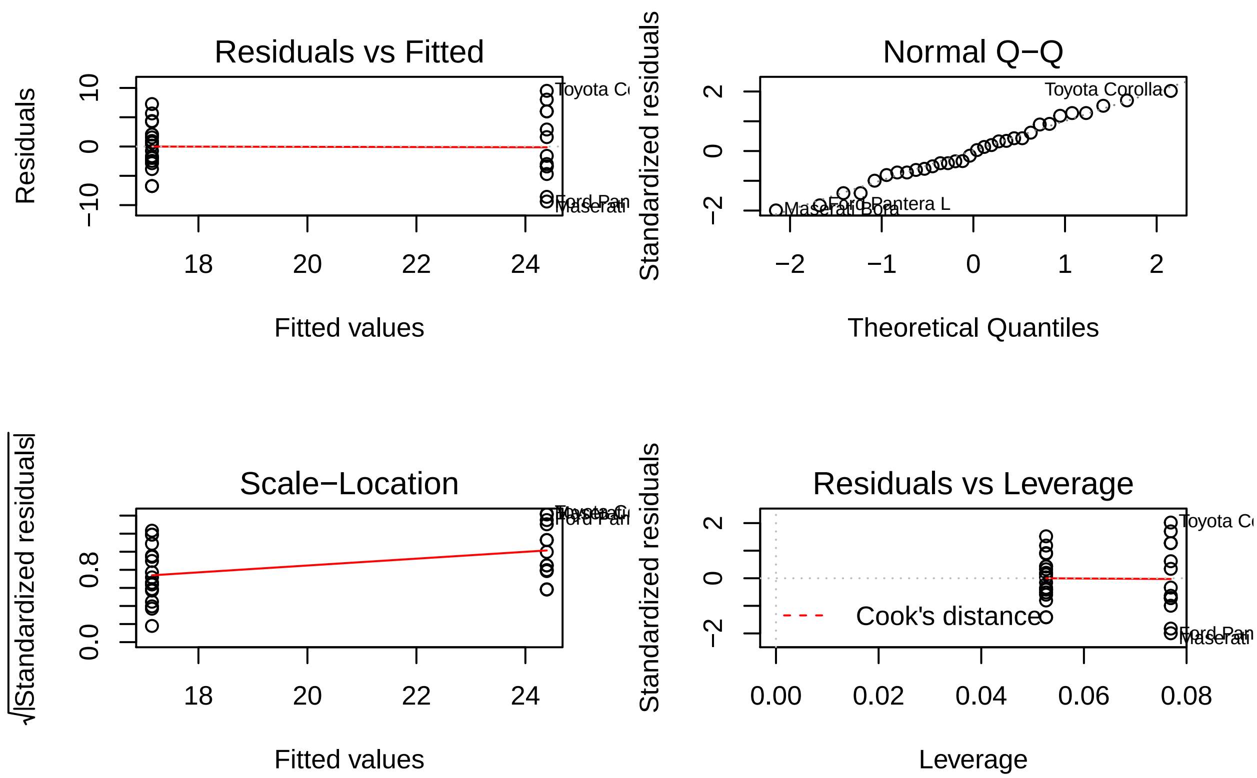
We see that our previous findings still hold. Driving manual has a positive impact in the mpg however we see that the effect dramastically decreased to 0.18 by adding more regressors to our model. The effect of the transmission on the MPG will change depending in which variables we choose to explain t he variability of our modell. After doing an analiysis of variance (ANOVA) and looking at the pvalues we see that the inclussion of these variables is indeed statistical significant for the model.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + wt + cyl
##   Res.Df    RSS Df Sum of Sq      F          Pr(>F)
## 1     30 720.90
## 2     28 191.05  2    529.85 38.828 0.000000008428 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is why it is important on having technical expertise to know which parameters to include in the model in order to choose the right model to fit the data.

Apendix and Extras

```r
par(mfrow = c(2,2))
plot(mdl)
```



Using the step function we can do model selection and find the best parameter combination to explain the variability of the model.

```r
mdl_best <- step(lm(mpg~.,data = mtcars), trace = 0)
summary(mdl_best)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382   0.177915
## wt           -3.9165     0.7112  -5.507 0.00000695 ***
## qsec          1.2259     0.2887   4.247   0.000216 ***
## am            2.9358     1.4109   2.081   0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 0.0000000000121
```

We see that the chosen variables for the model are actually wt, qsec and am. Most important is the R-squared for this model went up to 0.8496636 instead of 0.3597989 from the mdl with just the transmission as a parameter.